“Geocollective”: A Tool for Harvesting Geographic Information From the Social Web

Matthew Tenney[1], G. Brent Hall[3] and Renee Sieber[2]

1&2 Geography, McGill University,
Matthew.Tenney@mail.mcgill.ca, Renee.Sieber@mcgill.ca
3 Education and Research, Esri Canada
bhall@esri.ca

Abstract

Geocollective is a proof-of-concept project composed of largely open-source software; forming back-end tools required for the collection, processing, and basic scientific analysis of content collected from the Internet. Social media, mobile-technologies, and user-generated content of all kinds have piqued the interest of many social science research programs. Despite this interest there are still a great many challenges that range from use and licensing restrictions, programming environments and access points, as well as drifting domain expertise that separate experts based on technological skillsets and not interest or relevance to a particular phenomena.

Introduction

Geocollective is an attempt to bridge the gap between “traditional” and “new” forms of geographic information (GI) being generated on the Internet in academic scholarship (cf., Goodchild 1992; Coleman, Georgiadou, and Labonte 2009). The goal is to offer researchers an accessible means to gather and use relevant GI from these novel data sources regardless of their fields of inquiry or level of programming competency. While the need to develop new skills and techniques within human geography is a constant task, it is exacerbated by continual integration of social and technological worlds (Couclelis 1992; Elwood 2008; Sui 2008; Wilson 2012). Geocollective looks to extend the reach of the research community's “spatial perspective” to phenomena that can often transcend traditional concepts of spatial data and geographic analysis (cf., Gieryn 2000; Goodchild 2007). This is in part accomplished by providing tools like Geocollective that lower the demands on individual researchers and on available resources for acquiring disparate domain knowledge and tools such as the use of various programming languages and data service architectures. This helps to divert focus to the use and study of Web-GI (cf., Turner et al. 2008:1-13; Goodchild 2009; Jiang 2011; Brown et al. 2013).

Background and Relevance

Internet technologies have been engendered with numerous ideologies predating their actual creation (cf., Dahlberg 2001; DiMaggio and Garip 2012). However, there remains a sustained and common position that sees the Internet as an extension for society and social interactions (Papacharissi 2002; Good 2014). This extension provides a “space” for interaction between a few individuals to massive globally dispersed groups, as well as a “place” for the formation and participation of diverse communities of interest (Messinger et al. 2009; Zaphiris and Ang 2010:1–24). These characteristics of cyberspace have ignited disciplinary rattling discussions from the transcendence of distance (cf., Couclelis 1996; Wang, Lai, and Sui 2003; Tranos and Nijkamp 2013), time (cf., Loader and Dutton 2012; Lievrouw 2012; Karpf 2012), and social interaction (cf., Katz and Rice 2002; Katz 2007), all of which are fundamental dimensions responsible for driving geographic thought (Cresswell 2012).

The end of the 20th century witnessed a spread of the hyper-linked pages of the world-wide-web (Web). With the early years of the 21st century came a shift to user-produced ecosystem of ubiquitous content known as the Web 2.0 (Haklay, Singleton, and Parker 2008). Throughout the past 30-years geographers have kept a steady eye towards understanding, speculating, and integrating these evolving manifestations of a social reality by utilizing new technology and forms of (digital) data within our trade (e.g., Goodchild 2014). These changes become apparent as we consider things like the usability of geographic information systems (GIS) with the introduction of Google Earth (cf., Rana and Joliveau 2009), techniques for data collection with crowd-sourcing projects like OpenStreetMap (cf., Haklay and Weber 2008), and the conceptual impacts of amateur versus expert geographic information with the rise of “citizens as sensors” (cf., Goodchild 2008; Elwood, Goodchild, and Sui 2012).

Methods and Data

Built-in capabilities of Geocollective include combined application programming interfaces (API) that give users an ability to collect data easily from a variety of Web resources (cf., Abdalla 2012; Hu 2012). Working datasets can be obtained from several social media and content sharing services like Twitter, Facebook, Google+, YouTube and Flickr, as well as being saved into interoperable data models (e.g., delineated-text and Esri shapefiles) or spatial databases (e.g., Esri geodatabase and Postgres-PostGIS) according to the limitations of use for each particular service.

Geocollective is written in the high-level programming language of Python. This provides a customizable and readable code-base that can be integrated with a variety of popular front-end applications such as desktop GIS and web-interfaces. Additional data collection resources include spatial data from OpenStreetMap and content from Wikipedia, which add a robustness of crowd-based “collective knowledge” and supplemental structuring resources for working datasets (cf., Sigurbjörnsson and van Zwol 2008; Bizer et al. 2009; Mitra 2012). The ability to specify by geographic area, user, or key-word topics adds a search functionality to Geocollective for gathering data relevant to a user’s specific purpose.

Each API and data service has its own restrictions on the ability to collect or store content due to licensing specifications and download rates. Geocollective provides background services that follow these specifications and automate tasks for projects that require extended data-collection and observation periods (see Figure 1 for data harvesting work flow).

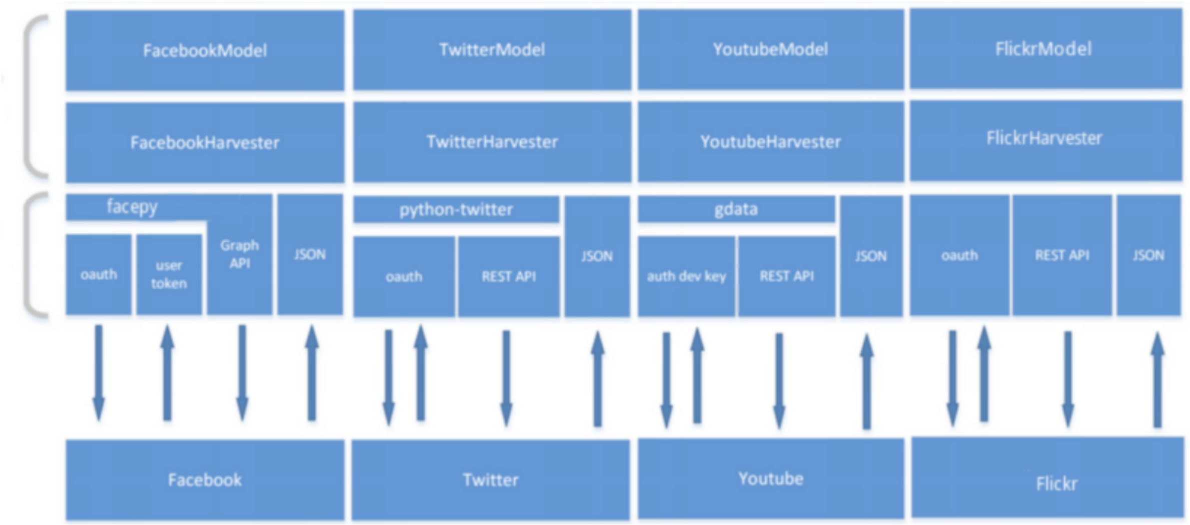

Figure 1: Example Web harvesting workflow for Geocollective

Geocollective includes several analytical techniques tailored specifically to the content it harvests, which have been replicated from contemporary literature. These techniques act as extensible tools for the analysis and application of the working datasets much like the role simple geometric functions (e.g., buffers, intersects) provide in traditional desktop and Web GIS.

The first set of tools focuses on the temporal and geographic qualities of the dataset (cp., Russell 2013). While most GIS platforms provide means to conduct spatial analysis there is often a need to consider Internet and user-generated data as time-series observations (Green 2002). More specifically, with the fast genesis of Web content the ability to translate timezones, query by timeliness, and aggregate by custom periods is often a fundamental need in many projects that make use of Web data.

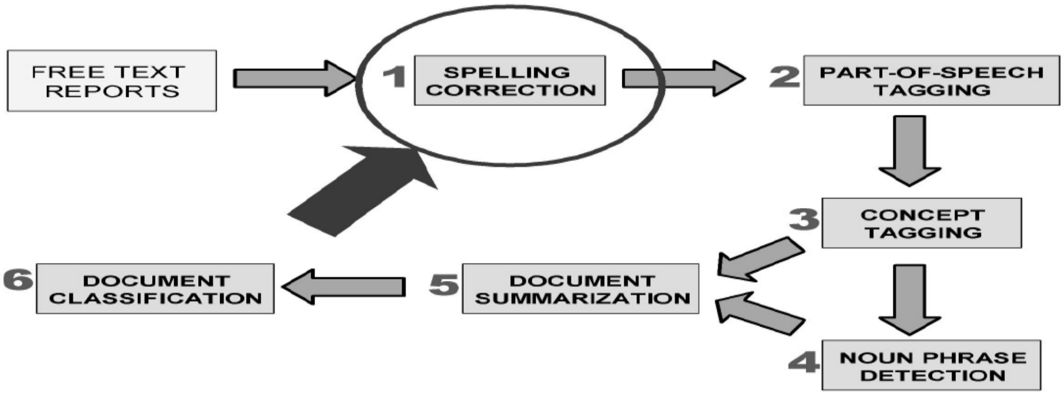

Figure 2: Example text processing workflow for Geocollective

Another key functionality included in Geocollective is basic computation linguistic techniques for dealing with unstructured text. These natural language processing (NLP) techniques let users normalize (cf., Pennell and Liu 2014) and extract more information from the free-form text (cf., Adams and McKenzie 2013; Vasardani, Winter, and Richter 2013). Several basic elements in the pre-processing stages are also included such as spell correction, word and character substitution or removal, position of speech tagging, as well as noun phrase and named entity extraction (see Figure 2 for text processing workflow).

Geocollective also provides several algorithms to calculate similarity measures between text content of working datasets and relating it to external resources like news-feeds, Wikipedia articles, and other corpora (cf., Ghosh and Guha 2013). Several experimental techniques for measuring text similarity are also being included like the semantic topic modeling "Word2Vec" algorithm developed by Google (Goldberg and Levy 2014), and extended by Le and Mikolov (2014), that has been shown to compare texts of varying lengths with good accuracy.

Geocollective is ultimately focused on collecting data dealing with human social interaction and inspection of the data itself as a socially created phenomenon (Wilson 2014). There is often an inherent property of social media in the form of a social network or graph structure (Johnson and Gilles 2003). An issue with graph representations of social dynamics is often attributed to an inability for capturing other influences such as space-time interactions (cf., Radil, Flint, and Tita 2010; Daraganova et al. 2012; adams, Faust, and Lovasi 2012; Takhteyev, Gruzd, and Wellman 2012; Stephens and Poorthuis 2014). Thus, there is some discussion in this presentation regarding the development of spatially and temporally situated social network graphs. That is, traditional social network graphs linked by social and topical relationships to geographic space by "content hubs" and interconnected through temporal connections.

Conclusions

This paper has presented a collaborative project called "Geocollective". The Geocollective project is in its first year of development for a proof-of-concept version, and some of the functionalities currently included and proposed for development were discussed.

A primary goal of Geocollective is to provide social scientists a means to gather easily data being generated from disparate Internet resources for the inclusion of their own research agendas. The involvement of specialists communities such as those found as SKI2015 are imperative to the growth and success of a project like Geocollective.

Techniques from basic text processing to semantic topic modeling were discussed in reference to their ability to tie further unstructured content and non-spatial observations to geographic locations through Geocollective's functionality. Additionally, novel data models like spatially and temporally situated social network graphs were proposed in light of their potential utility to contemporary geographic analysis as a means to possibly being more appropriate in representing dynamic human social networks and information flows across space and time.

Support and Acknowledgements

This research was supported by Esri Canada Inc., Geothink Canada, and the Mitacs Accelerate PhD Fellowship program.

References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 154–165. doi:10.1016/j.websem.2009.07.002

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research*, *4*(1), 332–358.

Couclelis, H. (1992). People manipulate objects (but cultivate fields): Beyond the raster-vector debate in GIS. In A. U. Frank, I. Campari, & U. Formentini (Eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space* (pp. 65–77). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/3-540-55966-3_3

Elwood, S. (2008). Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, *72*(3-4), 133–135. doi:10.1007/s10708-008-9187-z

Gieryn, T. F. (2000). A space for place in sociology. *Annual Review of Sociology*, 463–496.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722 [cs, Stat]*. Retrieved from http://arxiv.org/abs/1402.3722

Goodchild, M. F. (1992). Geographical information science. *International Journal of Geographical Information Systems*, *6*(1), 31–45. doi:10.1080/02693799208901893

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M. F. (2008). Commentary: whither VGI? *GeoJournal*, *72*(3-4), 239–244. doi:10.1007/s10708-008-9190-4

Goodchild, M. F. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services*, *3*(2), 82–96. doi:10.1080/17489720902950374

Goodchild, M. F. (2014). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science*, *0*(1), 3–20. doi:10.5311/josis.v0i1.32

Good, K. D. (2014). Internet, society and culture: Communicative practices before and after the Internet. *New Media & Society*, *16*(3), 536–537. doi:10.1177/1461444813518888a

Green, N. (2002). On the Move: Technology, Mobility, and the Mediation of Social Time and Space. *The Information Society*, *18*(4), 281–292. doi:10.1080/01972240290075129

Haklay, M., Singleton, A., & Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass*, *2*(6), 2011–2039. doi:10.1111/j.1749-8198.2008.00167.x

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12–18. doi:10.1109/MPRV.2008.80

Hu, S. (2012). Online Map Service Using Google Maps API and Other JavaScript Libraries: An Open Source Method. In M. P. Peterson (Ed.), *Online Maps with APIs and WebServices* (pp. 265–278). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-27485-5_17

Jiang, B. (2011). Making GIScience research more open access. *International Journal of Geographical Information Science*, *25*(8), 1217–1220. doi:10.1080/13658816.2011.585613

Johnson, C., & Gilles, R. P. (2003). Spatial social networks. In *Networks and Groups* (pp. 51–77). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24790-6_4

Karpf, D. (2012). Social Science Research Methods in Internet Time. *Information, Communication & Society*, *15*(5), 639–661. doi:10.1080/1369118X.2012.665468

Katz, J. E. (2007). Mobile Media and Communication: Some Important Questions. *Communication Monographs*, *74*(3), 389–394. doi:10.1080/03637750701543519

Katz, J. E., & Rice, R. E. (2002). Syntopia: Access, Civic Involvement, and Social Interaction on the Net. In B. Wellman & C. Haythornthwaite (Eds.), *The Internet in Everyday Life* (pp. 114–138). Blackwell Publishers Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470774298.ch3/summary

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*. Retrieved from http://arxiv.org/abs/1405.4053

Lievrouw, L. A. (2012). The Next Decade in Internet Time. *Information, Communication & Society*, *15*(5), 616–638. doi:10.1080/1369118X.2012.675691

Loader, B. D., & Dutton, W. H. (2012). A Decade in Internet Time. *Information, Communication & Society*, *15*(5), 609–615. doi:10.1080/1369118X.2012.677053

Messinger, P. R., Stroulia, E., Lyons, K., Bone, M., Niu, R. H., Smirnov, K., & Perelgut, S. (2009). Virtual worlds — past, present, and future: New directions in social computing. *Decision Support Systems*, *47*(3), 204–228. doi:10.1016/j.dss.2009.02.014

Mitra, A. (2012). Collective narrative expertise and the narbs of social media. In *Social Software and the Evolution of User Expertise: Future Trends in Knowledge Creation and Dissemination* (pp. 1–12).

Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society*, *4*(1), 9–27. doi:10.1177/14614440222226244

Pennell, D. L., & Liu, Y. (2014). Normalization of informal text. *Computer Speech & Language*, *28*(1), 256–277. doi:10.1016/j.csl.2013.07.001

Radil, S. M., Flint, C., & Tita, G. E. (2010). Spatializing Social Networks: Using Social Network Analysis to Investigate Geographies of Gang Rivalry, Territoriality, and

Violence in Los Angeles. *Annals of the Association of American Geographers*, *100*(2), 307–326. doi:10.1080/00045600903550428

Rana, S., & Joliveau, T. (2009). NeoGeography: an extension of mainstream geography for everyone made by everyone? *Journal of Location Based Services*, *3*(2), 75–81. doi:10.1080/17489720903146824

Russell, M. A. (2013). *Mining the social web*.

Sigurbjörnsson, B., & van Zwol, R. (2008). Flickr Tag Recommendation Based on Collective Knowledge. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 327–336). New York, NY, USA: ACM. doi:10.1145/1367497.1367542

Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, *32*(1), 1–5. doi:10.1016/j.compenvurbsys.2007.12.001

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81. doi:10.1016/j.socnet.2011.05.006

Tranos, E., & Nijkamp, P. (2013). The Death of Distance Revisited: Cyber-Place, Physical and Relational Proximities. *Journal of Regional Science*, *53*(5), 855–873. doi:10.1111/jors.12021

Turner, A., Forrest, B., Lorica, B., & Magoulas, R. (2008). *Where 2.0: The State of the Geospatial Web*. O'Reilly Media, Incorporated. Retrieved from http://shop.oreilly.com/product/9780596522568.do

Vasardani, M., Winter, S., & Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, *27*(12), 2509–2532. doi:10.1080/13658816.2013.785550

Wang, Y., Lai, P., & Sui, D. (2003). Mapping the Internet using GIS: The death of distance hypothesis revisited. *Journal of Geographical Systems*, *5*(4), 381–405. doi:10.1007/s10109-003-0117-9

Wilson, M. W. (2012). Location-based services, conspicuous mobility, and the location-aware future. *Geoforum*, *43*(6), 1266–1275. doi:10.1016/j.geoforum.2012.03.014

Wilson, M. W. (2014). Morgan Freeman is dead and other big data stories. *Cultural Geographies*, 1474474014525055. doi:10.1177/1474474014525055

Zaphiris, P., & Ang, C. S. (2010). *Social computing and virtual communities*. Boca Raton: Chapman & Hall/CRC Press.